elixir ITALY

DAY 4

Discovery of genomic structural variations using NGS data

26.11.2015

Pille Hallast pille.hallast@ut.ee

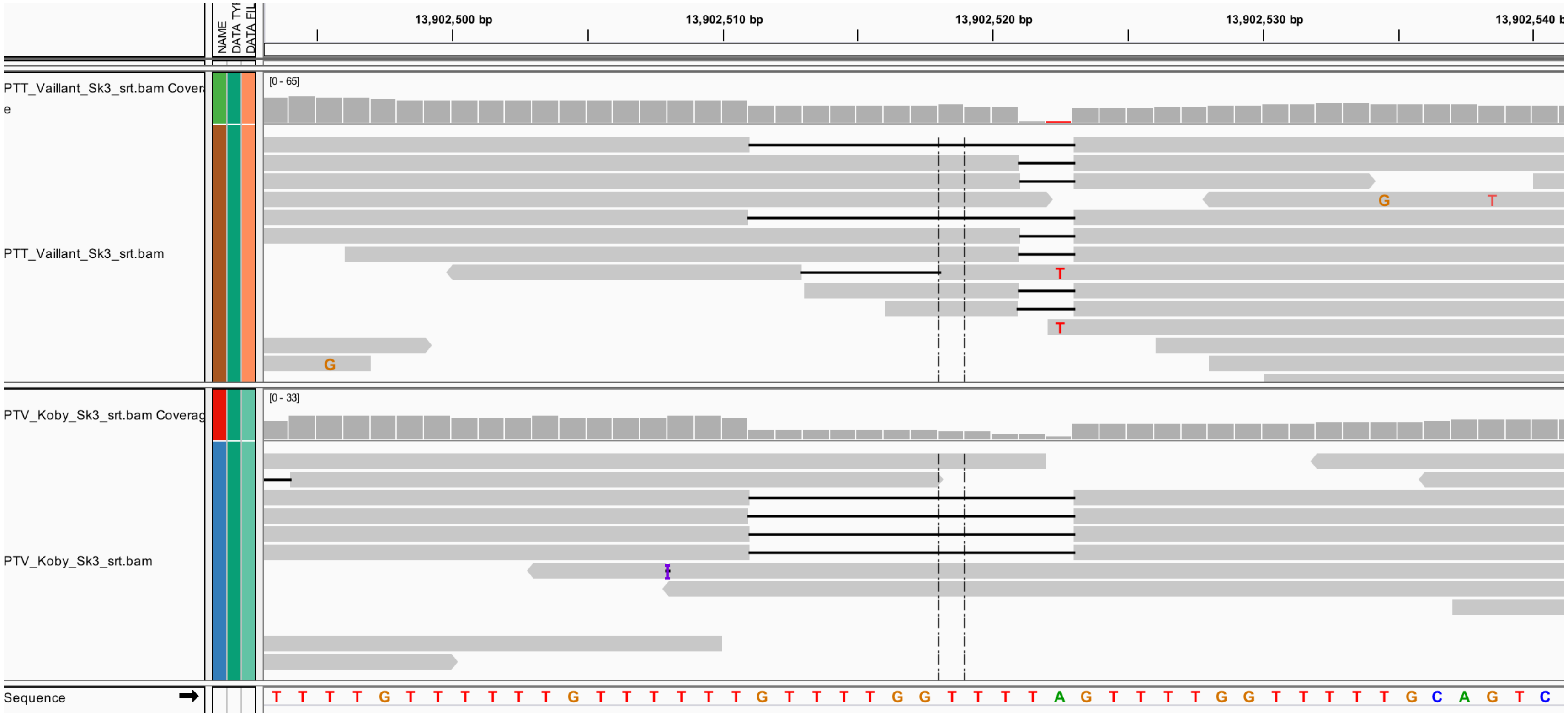
UNIVERSITY OF TARTU 1632

Overview

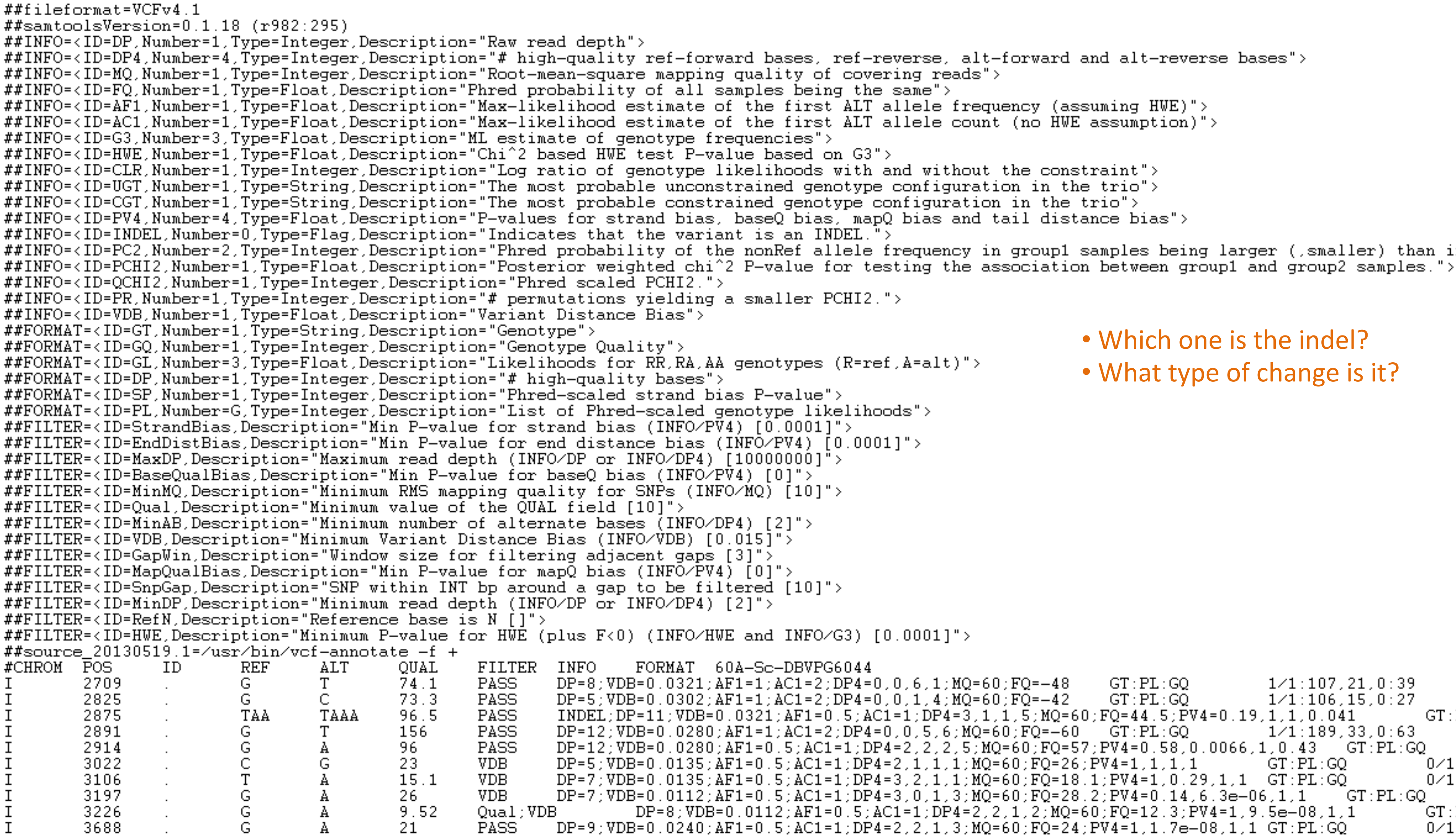
- Small indels
- Genomic rearrangements/structural variations (SV)
- Discovery of SVs
  - Other methods used so far
  - Using NGS data:
    - Whole genomes
    - Capture data (e.g. exomes, custom capture)
    - other
- Practical using software pindel

Small indels

- 1 bp to a few tens of bp (in this context)
- Calling reliability depends on sequencing technology and type of data
- Most variant callers include small indel calling
- More care is needed (e.g. overlap between different callers)
- Validation!!

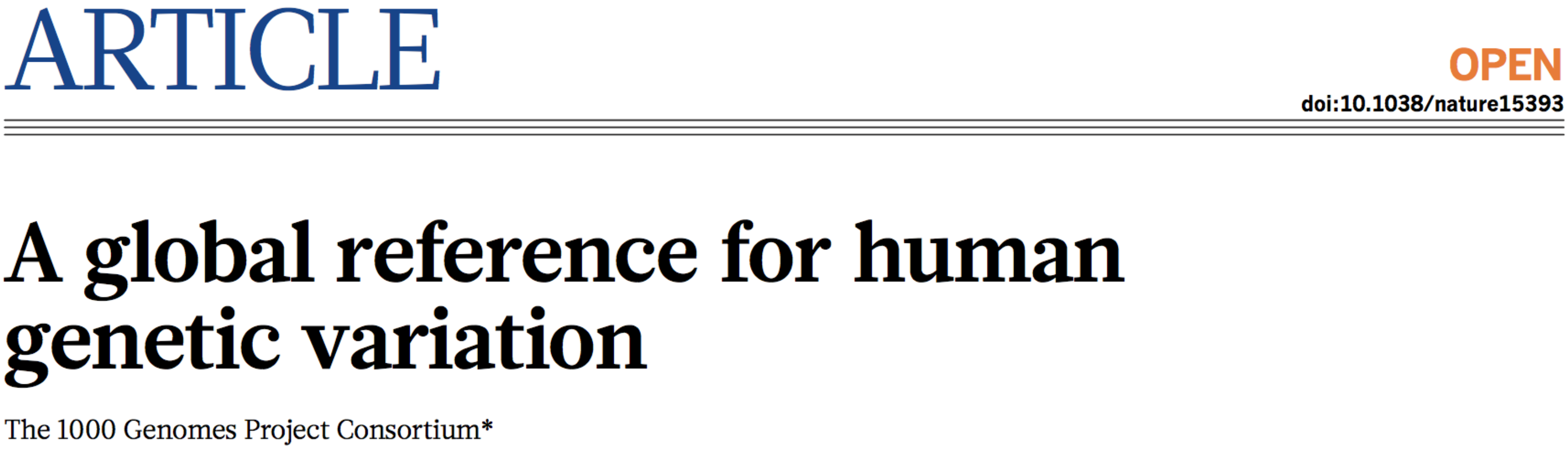


variant calling – VCF file



```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.18 (r982:295)
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype likelihoods with and without the constraint">
##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstrained genotype configuration in the trio">
##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrained genotype configuration in the trio">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than i
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FILTER=<ID=StrandBias,Description="Min P-value for strand bias (INFO/PV4) [0.0001]">
##FILTER=<ID=EndDistBias,Description="Min P-value for end distance bias (INFO/PV4) [0.0001]">
##FILTER=<ID=MaxDP,Description="Maximum read depth (INFO/DP or INFO/DP4) [10000000]">
##FILTER=<ID=BaseQualBias,Description="Min P-value for baseQ bias (INFO/PV4) [0]">
##FILTER=<ID=MinMQ,Description="Minimum RMS mapping quality for SNPs (INFO/MQ) [10]">
##FILTER=<ID=Qual,Description="Minimum value of the QUAL field [10]">
##FILTER=<ID=MinAB,Description="Minimum number of alternate bases (INFO/DP4) [2]">
##FILTER=<ID=VDB,Description="Minimum Variant Distance Bias (INFO/VDB) [0.015]">
##FILTER=<ID=GapWin,Description="Window size for filtering adjacent gaps [3]">
##FILTER=<ID=MapQualBias,Description="Min P-value for mapQ bias (INFO/PV4) [0]">
##FILTER=<ID=SnpGap,Description="SNP within INT bp around a gap to be filtered [10]">
##FILTER=<ID=MinDP,Description="Minimum read depth (INFO/DP or INFO/DP4) [2]">
##FILTER=<ID=RefN,Description="Reference base is N []">
##FILTER=<ID=HWE,Description="Minimum P-value for HWE (plus F<0) (INFO/HWE and INFO/G3) [0.0001]">
##source_20130519.1=/usr/bin/vcf-annotate -f +
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  60A-Sc-DBVPG6044
I       2709    .       G       T       74.1    PASS    DP=8;VDB=0.0321;AF1=1;AC1=2;DP4=0,0,6,1;MQ=60;FQ=-48    GT:PL:GQ        1/1:107,21,0:39
I       2825    .       G       C       73.3    PASS    DP=5;VDB=0.0302;AF1=1;AC1=2;DP4=0,0,1,4;MQ=60;FQ=-42    GT:PL:GQ        1/1:106,15,0:27
I       2875    .       TAA     TAAA    96.5    PASS    INDEL;DP=11;VDB=0.0321;AF1=0.5;AC1=1;DP4=3,1,1,5;MQ=60;FQ=44.5;PV4=0.19,1,1,0.041       GT:
I       2891    .       G       T       156     PASS    DP=12;VDB=0.0280;AF1=1;AC1=2;DP4=0,0,5,6;MQ=60;FQ=-60   GT:PL:GQ        1/1:189,33,0:63
I       2914    .       G       A       96      PASS    DP=12;VDB=0.0280;AF1=0.5;AC1=1;DP4=2,2,2,5;MQ=60;FQ=57;PV4=0.58,0.0066,1,0.43   GT:PL:GQ
I       3022    .       C       G       23      VDB     DP=5;VDB=0.0135;AF1=0.5;AC1=1;DP4=2,1,1,1;MQ=60;FQ=26;PV4=1,1,1,1       GT:PL:GQ        0/1
I       3106    .       T       A       15.1    VDB     DP=7;VDB=0.0135;AF1=0.5;AC1=1;DP4=3,2,1,1;MQ=60;FQ=18.1;PV4=1,0.29,1,1  GT:PL:GQ        0/1
I       3197    .       G       A       26      VDB     DP=7;VDB=0.0112;AF1=0.5;AC1=1;DP4=3,0,1,3;MQ=60;FQ=28.2;PV4=0.14,6.3e-06,1,1    GT:PL:GQ
I       3226    .       G       A       9.52    Qual;VDB        DP=8;VDB=0.0112;AF1=0.5;AC1=1;DP4=2,2,1,2;MQ=60;FQ=12.3;PV4=1,9.5e-08,1,1       GT:
I       3688    .       G       A       21      PASS    DP=9;VDB=0.0240;AF1=0.5;AC1=1;DP4=2,2,1,3;MQ=60;FQ=24;PV4=1,1.7e-08,1,1 GT:PL:GQ        0/1
```

- Which one is the indel?
- What type of change is it?

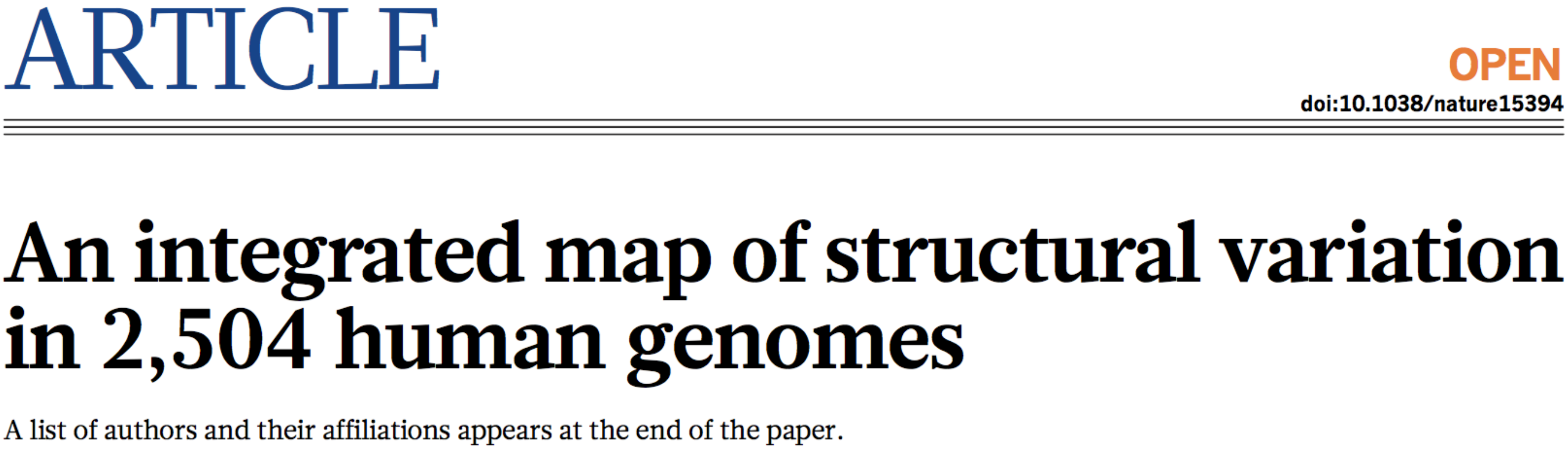


ARTICLE

OPEN
doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*



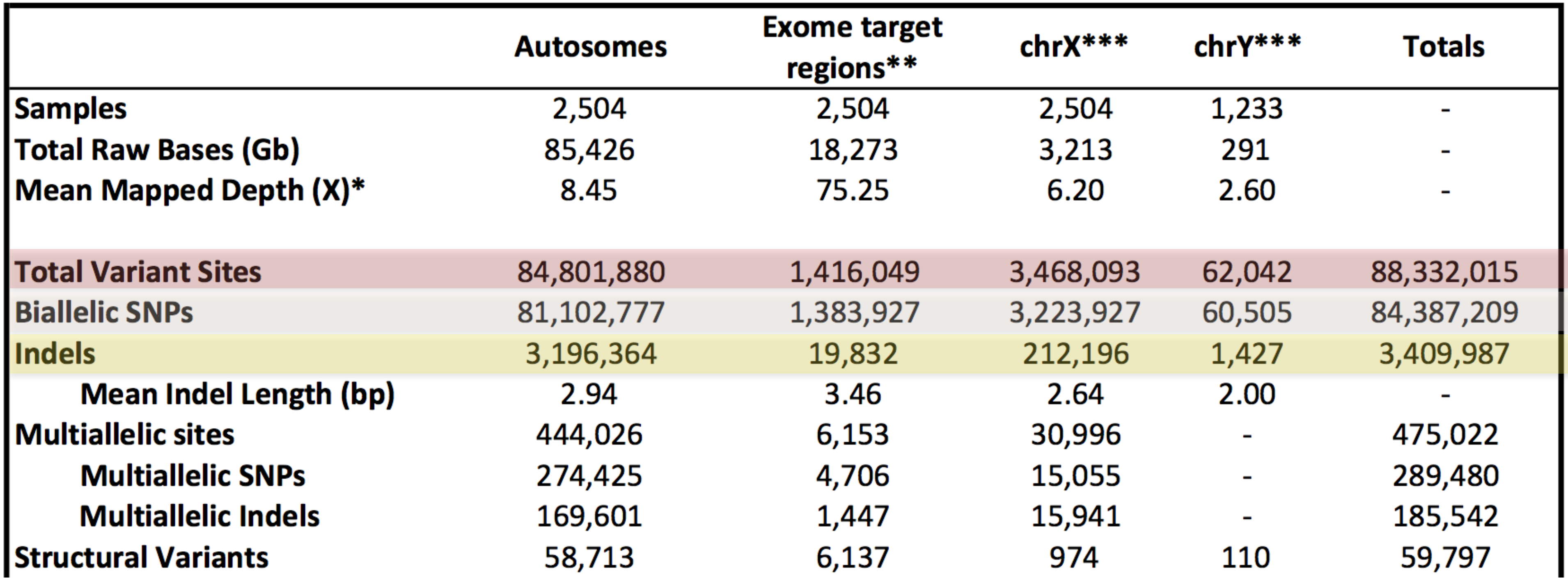
ARTICLE

OPEN
doi:10.1038/nature15394

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Nature 1 October 2015



	Autosomes	Exome target regions**	chrX***	chrY***	Totals
Samples	2,504	2,504	2,504	1,233	-
Total Raw Bases (Gb)	85,426	18,273	3,213	291	-
Mean Mapped Depth (X)*	8.45	75.25	6.20	2.60	-
Total Variant Sites	84,801,880	1,416,049	3,468,093	62,042	88,332,015
Biallelic SNPs	81,102,777	1,383,927	3,223,927	60,505	84,387,209
Indels	3,196,364	19,832	212,196	1,427	3,409,987
Mean Indel Length (bp)	2.94	3.46	2.64	2.00	-
Multiallelic sites	444,026	6,153	30,996	-	475,022
Multiallelic SNPs	274,425	4,706	15,055	-	289,480
Multiallelic Indels	169,601	1,447	15,941	-	185,542
Structural Variants	58,713	6,137	974	110	59,797

From 1000G phase 3 (Nature 2015 Oct 1)

Structural variations (SVs)

- Medium events (>10 bp up to 1 kb)
- Large events (typically >1kb to many Mb)
- Affect more bases than SNPs/short indels
  - 4.1 – 5M variant sites compared to the reference:
    - >99.9% - SNPs and short indels
    - 2,100 – 2,500 structural variants – affect ~20Mb!!

What are they?

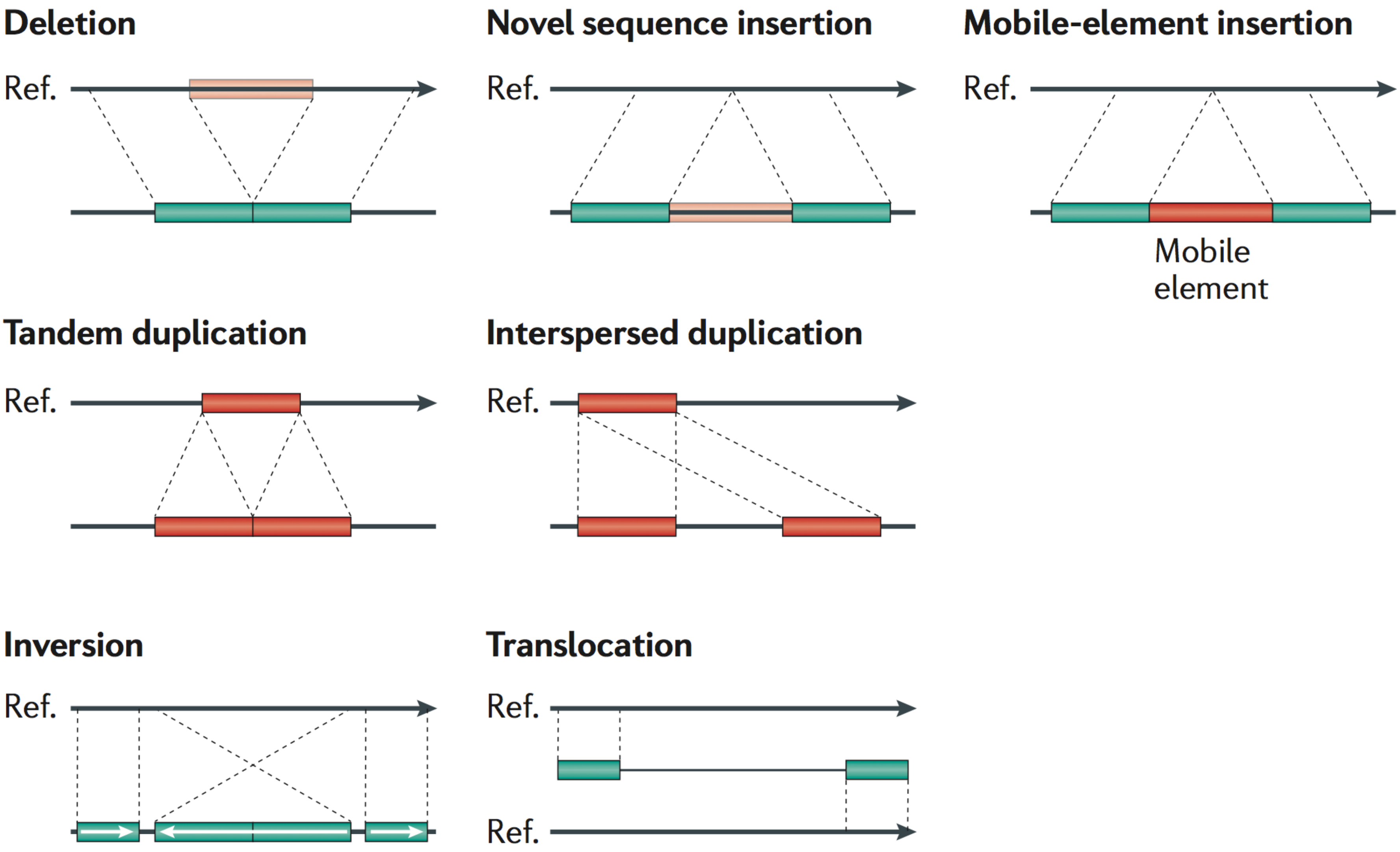


Figure from Nature Reviews Genetics 12, 363-376

Why study?

- Can have severe functional consequences as large genomic regions are affected
- Gene dosage/ gene disruption/ position/ unmasking a recessive allele etc
  - Charcot–Marie–Tooth disease type 1A – 1.4Mb duplication
  - hereditary neuropathy – 1.4Mb deletion
  - Smith–Magenis microdeletion syndrome – 3.7Mb deletion

    17p11.2p12
  - Sex-reversal – *SRY* chrY to chrX translocation
  - hemophilia A – inversion in factor VIII gene
  - Fragile X syndrome and Huntington's disease – trinucleotide repeat expansion

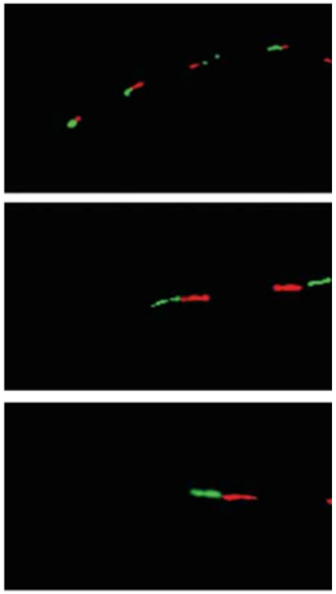
  - salivary amylase (AMY1) – digestion of starch, 1-10 copies

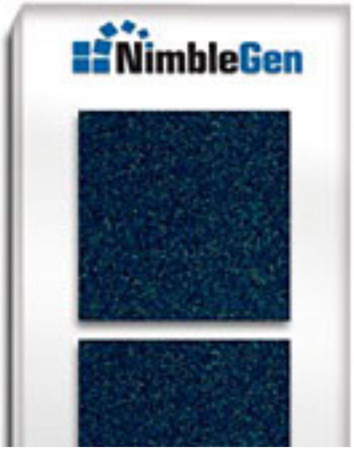
Methods used so far

- Fluorescent in situ hybridisation (FISH)
  - Fluorescent probes (~100kb)
  - Detect presence or absence of a specific region
- Array Comparative Genomic Hybridi
  - Test vs reference sample
  - Millions of probes, defining the resolution

ultra-high resolution 24-42M probes)

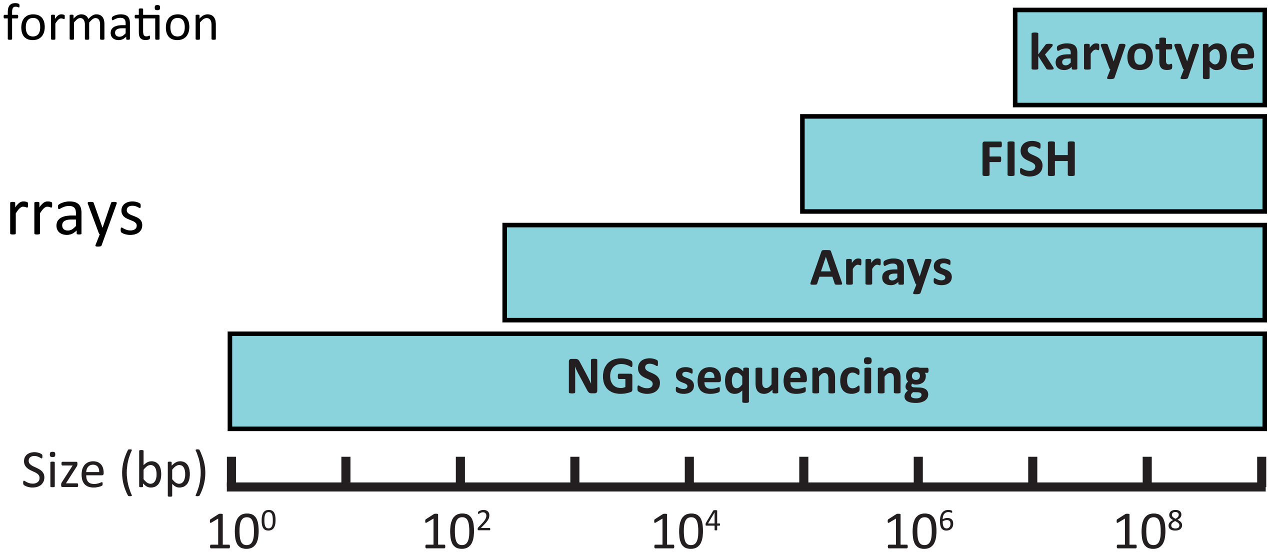
- SNP arrays
  - Single sample per microarray
  - Millions of probes, defining the resolution
  - Offer SNP genotype information





- Limitations of arrays
  - Dependent on the reference sequence used to design the probes
  - Lower resolution
  - No information about location
  - Cannot detect balanced events (e.g. inversions)
  - Performance in repeat-rich and duplicated regions is not great
  - No breakpoint information

- Advantages of arrays
  - Cheaper
  - Faster
  - High throughput



- Difficulties created by NGS data:
  - Quality of indel calls depend on the technology used
  - Relatively short read-lengths
  - Insert size
  - Sequence coverage

- Advantages of NGS data:
  - Possibility to detect different variants in a single experiment
  - Largely unbiased
  - Complete spectrum of genetic variation
  - Breakpoint information

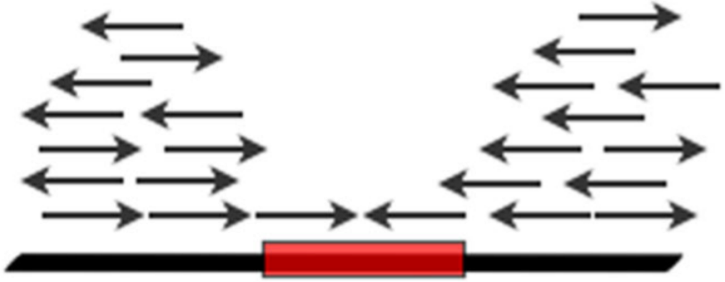
NGS data: methods

- Read depth
- Read-pair
- Split-read
- *De novo* assembly
- Combination of the above

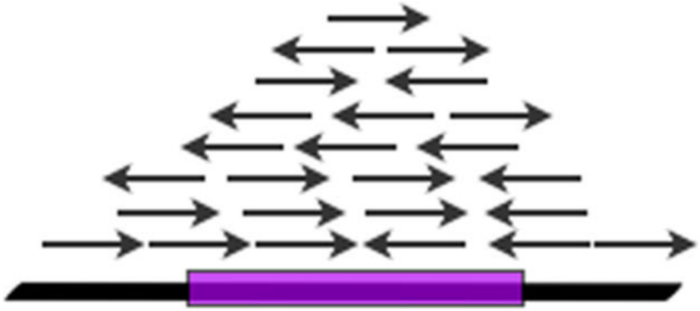
WGS data

- Read depth methods:
  - Assess read depth (higher DP – higher copy number)
  - Assume random (typically Poisson or modified Poisson) distribution of read depth
  - Study design for normalisation: single-sample; paired case/control; large population sample
  - Divergence from the distribution to identify:
    - Deletions – significantly reduced DP
    - Duplications – significantly higher DP
  - Absolute copy-number prediction
  - Usually poor break-point resolution
  - E.g. ReadDepth, mrCaNaVar, RDXplorer, CNV-seq, cn.MOPS and CNVnator

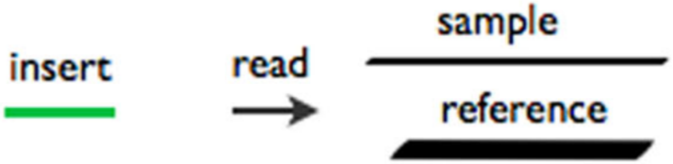
WGS data: read depth methods



Deletion



Tandem duplication



Modified from Tattini et al. 2015

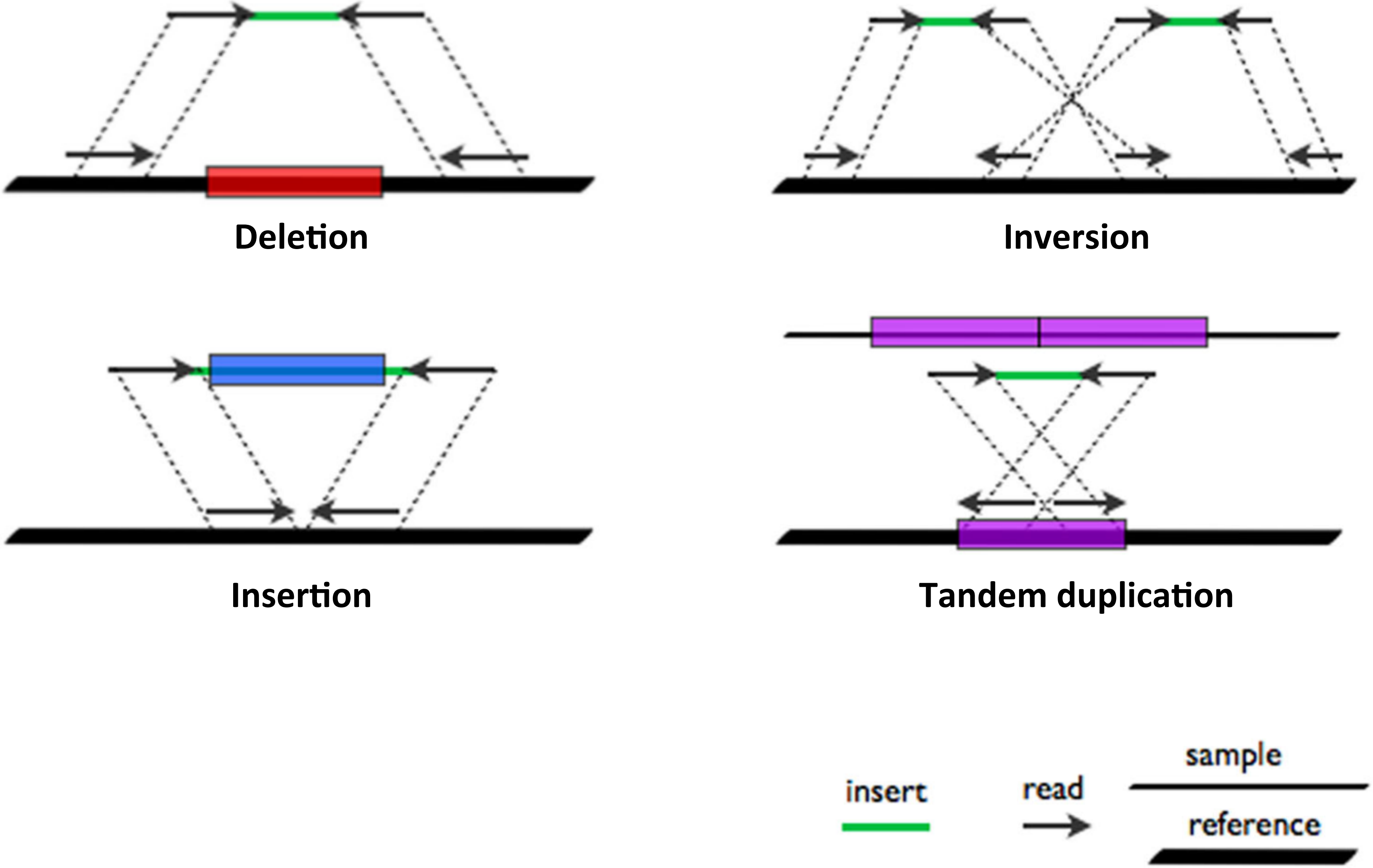
WGS data

- Read-pair methods:
  - Assess the span and orientation of paired-end reads
  - Require tight insert size distribution
  - Find discordant pairs where mapping span and/or orientation is discordant with reference genome
    - Deletions – mapping too far
    - Insertions – mapping too close
    - Inversions and some tandem duplications – orientation inconsistency

  - Can identify a wide range of rearrangements
  - Do not perform well in repetitive regions
  - Breakpoint prediction depends on very tight fragment size distribution

  - E.g. PEMer, VariationHunter, BreakDancer, MoDIL, MoGUL, HYDRA, Corona and SPANNER

WGS data: read-pair methods



Deletion

Inversion

Insertion

Tandem duplication

Modified from Tattini et al. 2015

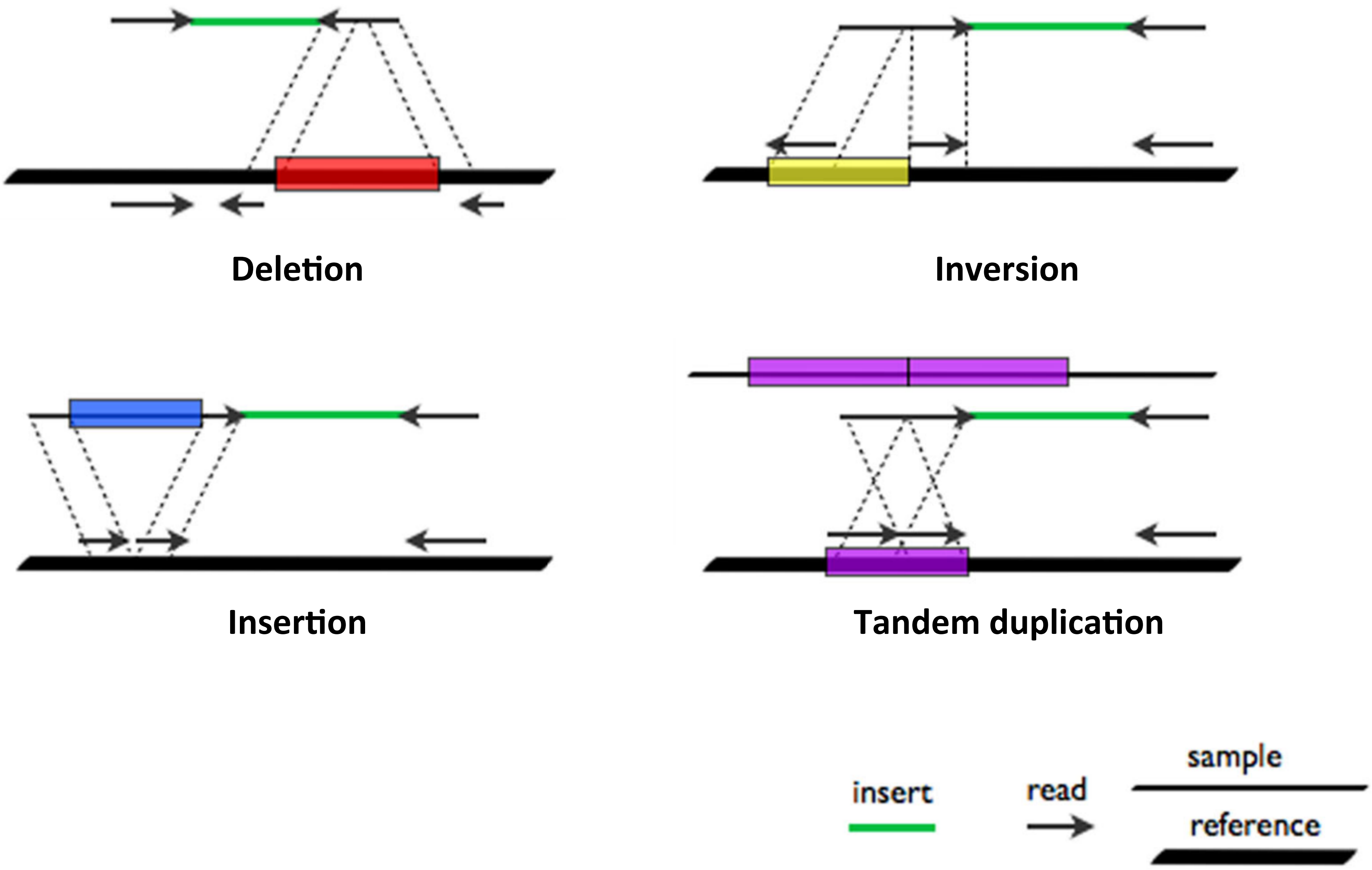
WGS data

- Split-read methods:
  - Identify the breakpoint of a SV based on a “split” in the read signature
  - One read maps uniquely, while the other fails to map or maps partly
  - Detection of a number of rearrangements
    - May be able to detect mobile-element insertions

  - Exact breakpoint resolution
  - Relies heavily on read length, but also insert size
  - Reliable only in unique parts of the genome

  - E.g. Pindel, AGE, Splitread, SLOPE, SRiC

WGS data: split-read methods



Deletion

Inversion

Insertion

Tandem duplication

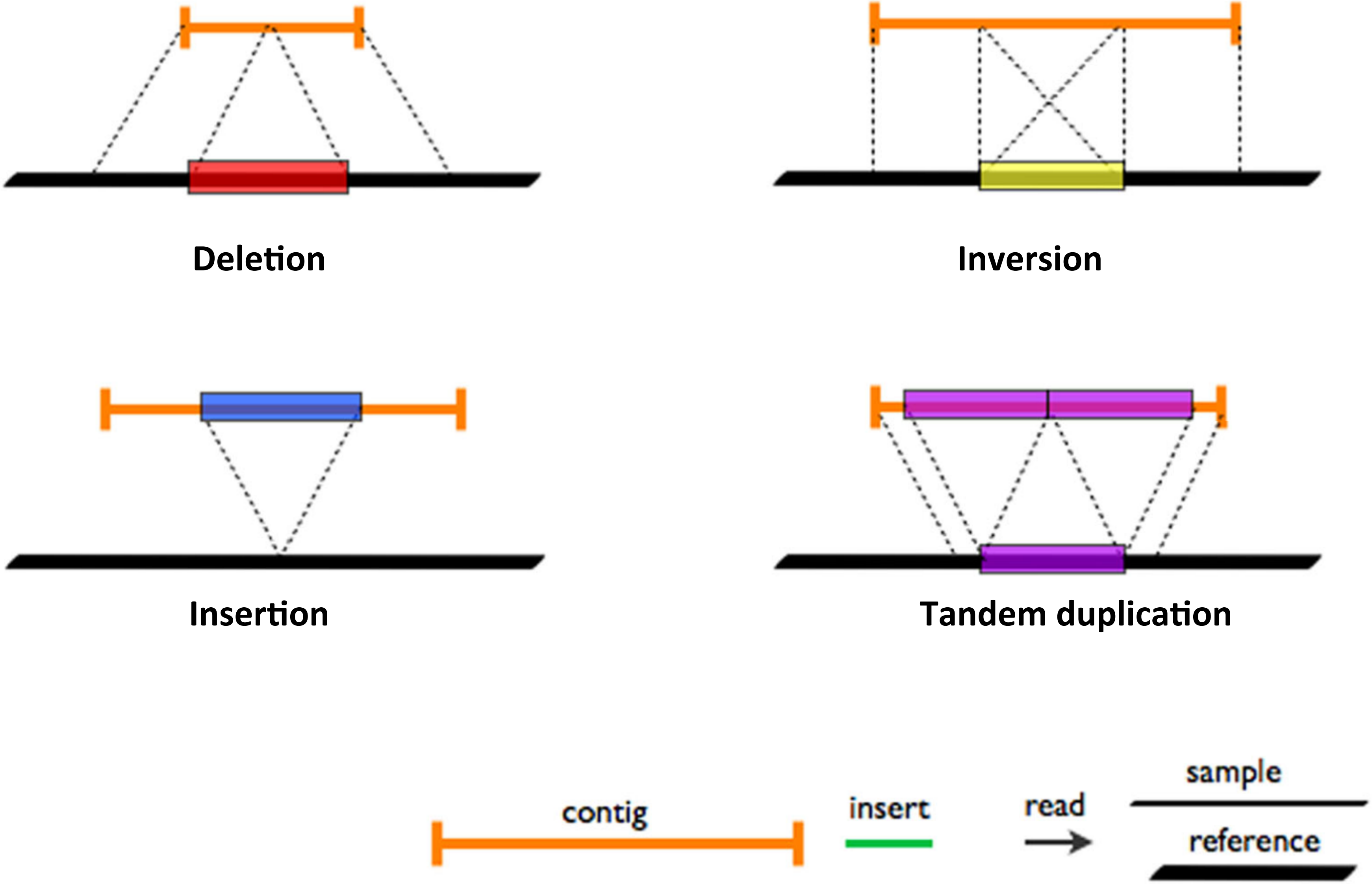
Modified from Tattini et al. 2015

WG-sequence data

- Sequence assembly:
  - *De novo* assembly of the sequence followed by comparison to a reference genome
    - should allow accurate typing of copy, content and structure
  - Often use combination of *de novo* and local assembly algorithms

  - Most promising of all the methods
  - Biased against repeats and duplications

  - E.g.
    - *de novo* assembly: EULER-USR, ABySS, SOAPdenovo and ALLPATHS-LG
    - Mixed approaches: Cortex, NovelSeq, TIGRA



Deletion

Inversion

Insertion

Tandem duplication

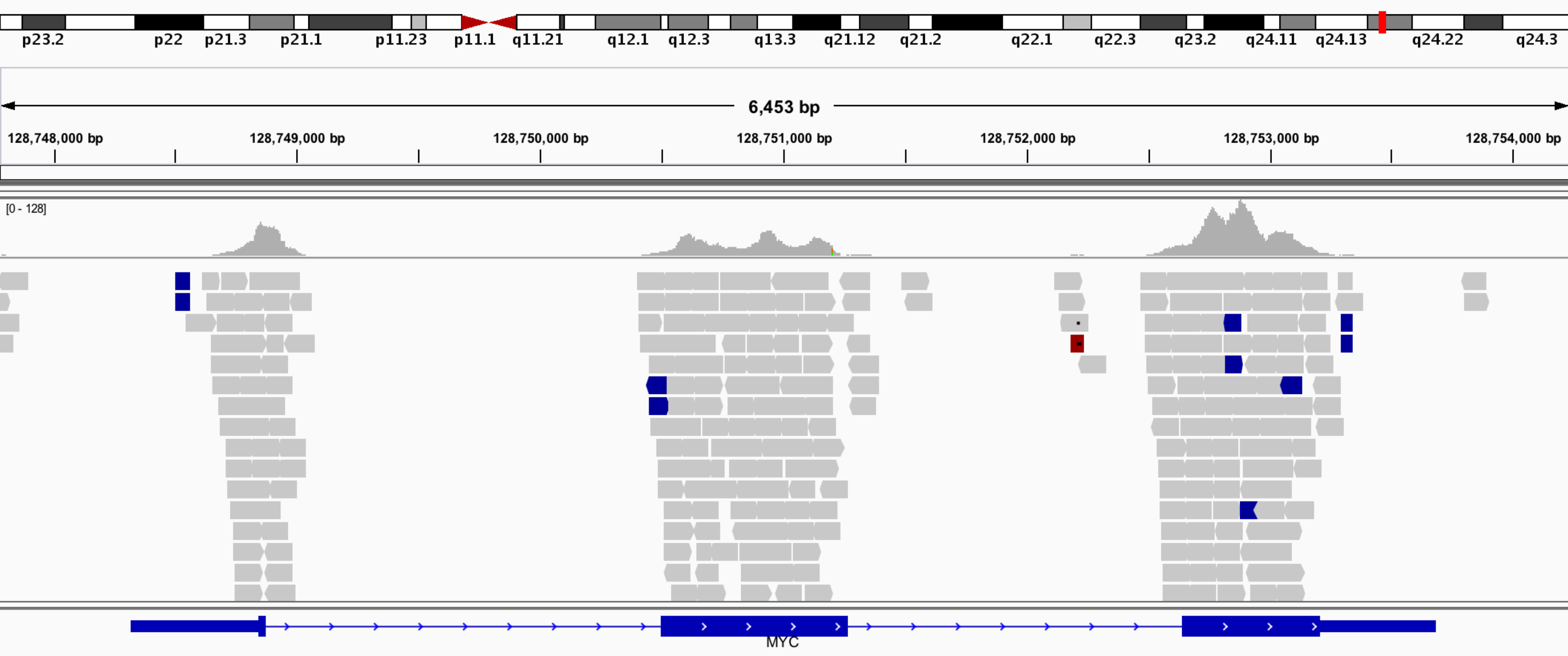
Modified from Tattini et al. 2015

Exome or custom capture data

- Characteristics of the data
  - Uneven coverage, but usually much higher depth
  - Inconsistent capture efficiency
  - Most SV breakpoints cannot be detected
  - More susceptible to GC-bias

  - Paired-end and split-read methods generally not well suited
  - Read depth methods
    - **Depth normalisation!!!**

  - E.g. CoNIFER, XHMM, ExomeCNV, VarScan2, SeqGene


p23.2
p22
p21.3
p21.1
p11.23
p11.1
q11.21
q12.1
q12.3
q13.3
q21.12
q21.2
q22.1
q22.3
q23.2
q24.11
q24.13
q24.22
q24.3
6,453 bp
128,748,000 bp
128,749,000 bp
128,750,000 bp
128,751,000 bp
128,752,000 bp
128,753,000 bp
128,754,000 bp
[0 - 128]
MYC

Exome or custom capture data

- Characteristics of the data
  - Uneven coverage, but usually much higher depth
  - Inconsistent capture efficiency
  - Most SV breakpoints cannot be detected
  - More susceptible to GC-bias

  - Paired-end and split-read methods generally not well suited
  - Read depth methods
    - **Depth normalisation!!!**

  - E.g. CoNIFER, XHMM, ExomeCNV, VarScan2, SeqGene

Amplicon sequencing

- Characteristics of the data
  - Different biases compared to WGS or WES data
  - Normalisation less effective due to limited target regions
  - High depth but very heterogeneous
  - GC-bias and amplicon length bias
  - Read depth methods
    - **Depth normalisation!!!**
  - E.g. ONCOCNV, AMS

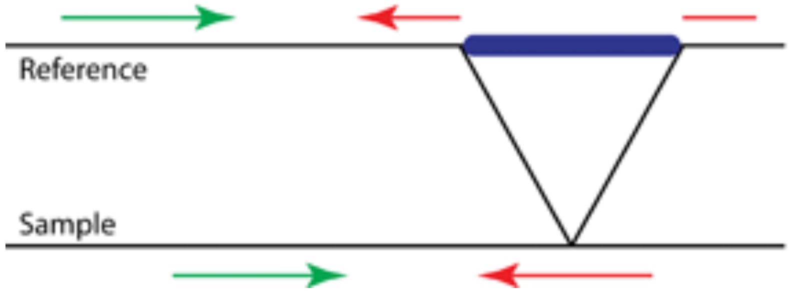
Practical:

1. Obtain the insert size metrics for the samples using picard
2. Use Pindel to call a range of SVs
3. Convertion of pindel calls into vcf with some filtering using pindel2vcf

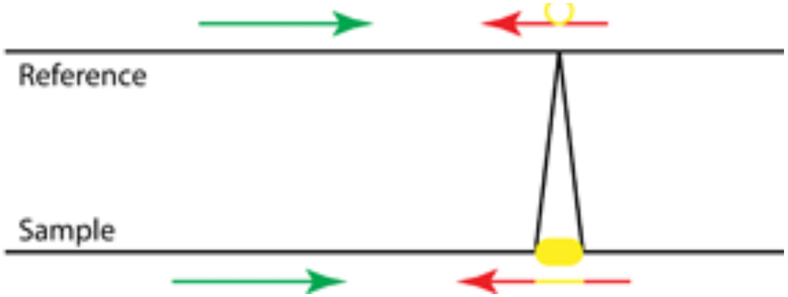
Pindel - A pattern growth approach, split-read method

Simple events:

Deletion

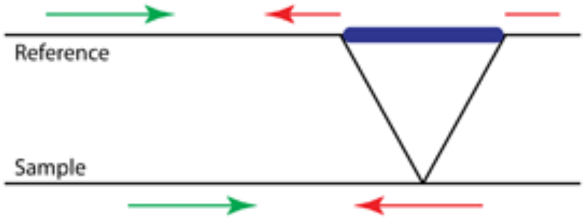


Insertion

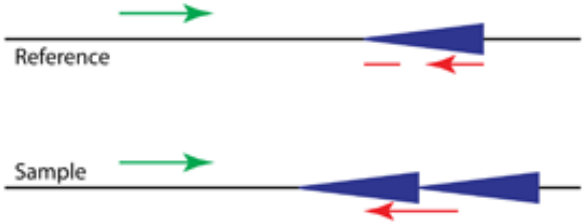


Complex events:

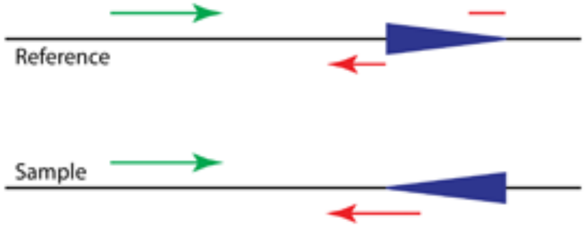
Large deletion



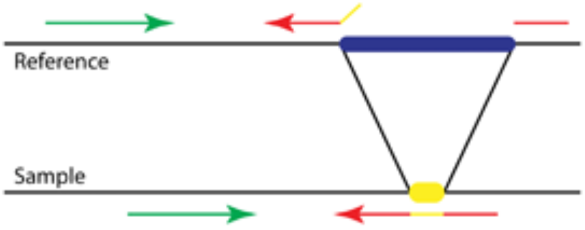
Tandem duplication

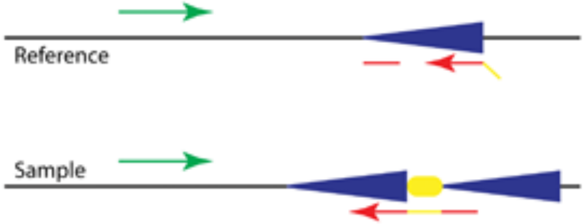


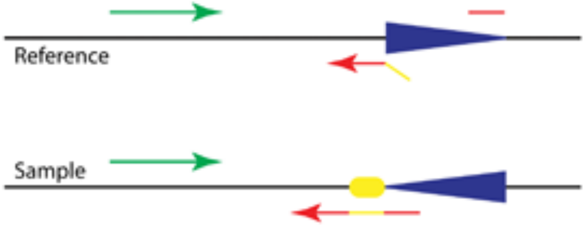
Inversion



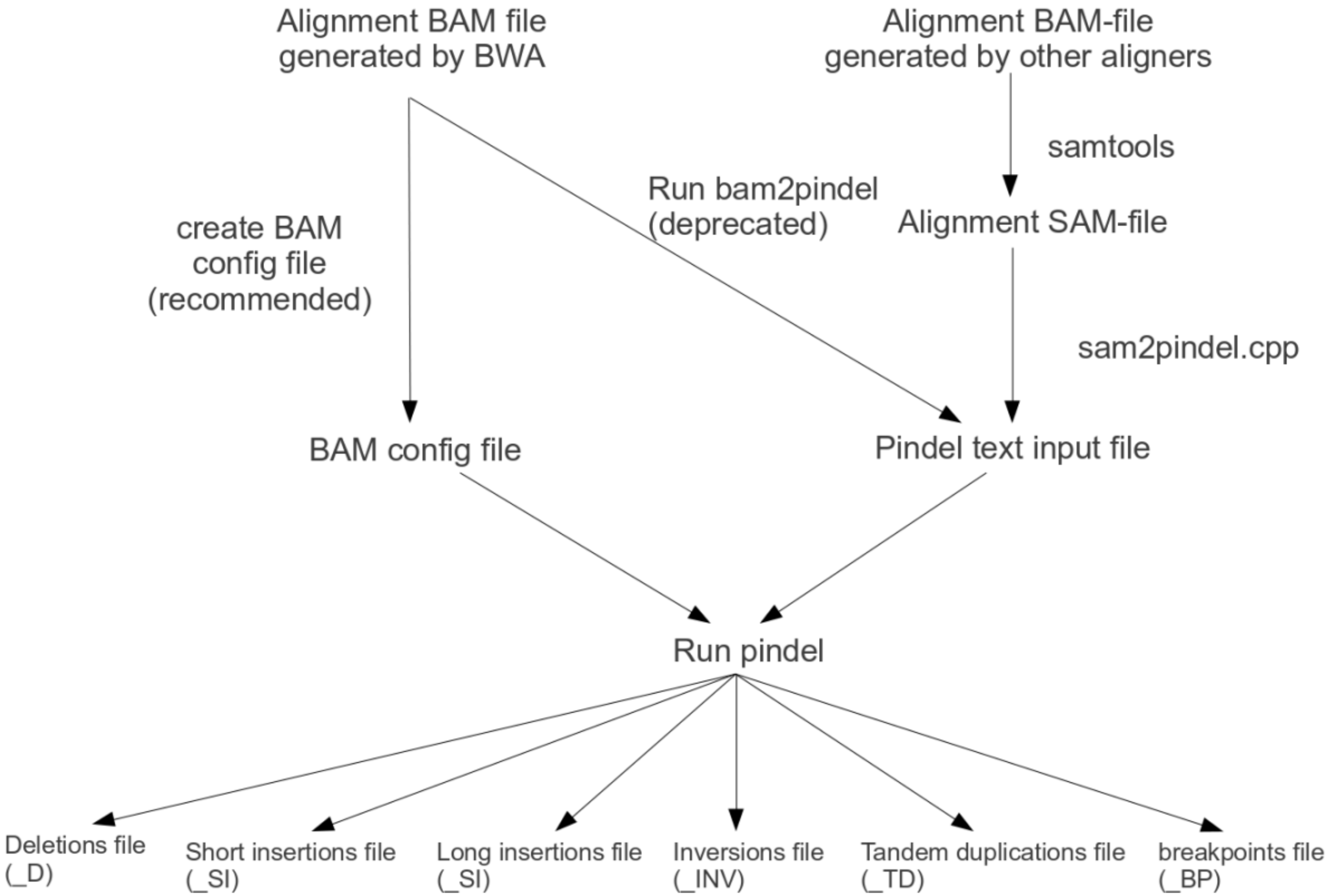
With additional sequence







Pindel workflow:



Alignment BAM file generated by BWA

Alignment BAM-file generated by other aligners

samtools

Run bam2pindel (deprecated)

Alignment SAM-file

create BAM config file (recommended)

sam2pindel.cpp

BAM config file

Pindel text input file

Run pindel

Deletions file (_D)

Short insertions file (_SI)

Long insertions file (_SI)

Inversions file (_INV)

Tandem duplications file (_TD)

breakpoints file (_BP)

Make a new folder and copy the data:

```
cd /pico/scratch/usertrain/your_username/
mkdir SV
cd SV

cp /pico/scratch/userexternal/phallast/SV/bams/*.bam .
cp /pico/scratch/userexternal/phallast/SV/bams/*.bai .
cp /pico/scratch/userexternal/phallast/project2/ref/ucsc.hg19.fasta .
cp /pico/scratch/userexternal/phallast/project2/ref/ucsc.hg19.dict .
cp /pico/scratch/userexternal/phallast/project2/ref/ucsc.hg19.fasta.fai .
```

The folder contains:

- bam files from two whole-genome sequenced chimpanzees Vaillant and Koby, sequenced on HiSeq 2000, with 100bp paired-end reads
- The bam files contain ~2.2Mb region of the Y chromosome (chrY: 13870437-16095787)
- Raw data has already been mapped to human genome reference hg19 using bwa mem v0.7.1 followed by standard bam refinement

- From Prado-Martinez et al "Great ape genetic diversity and population history." Nature. 2013 Jul 25;499(7459):471-5

Obtain the insert size metrics for the samples using picard

```
module load profile/advanced
module load autoload picard/1.119

java -Xmx1G -jar /cineca/prod/applications/picard/1.119/binary/
bin/CollectInsertSizeMetrics.jar I=PTV_Koby_Sk3_srt.bam
O=PTV_Koby_Sk3_srt_InsMetr.txt H=PTV_Koby_Sk3_srt_hist.pdf

ls -l
more PTV_Koby_Sk3_srt_InsMetr.txt
```

Run the same for the other sample.

What is the mean insert size for these two samples?

Parameters used to run picard:

I	Input bam file
O	Output file containing Insert size metrics
H	insert_size_Histogram.pdf

Create a configuration file for pindel

Use a text editor to chreate a text file in the following format (separated by tabs):
on each line, list the name of the bam-file, the insert size, and the label for the sample.
For example:

```
path-to-data/sample1.bam	236	sample1
path-to-data/sample2.bam	324	sample2
```

Run Pindel:

```
module load profile/advanced
module load autoload pindel/1.0
pindel -h

pindel -f /pico/scratch/usertrain/your_username/SV/
ucsc.hg19.fasta -i conf_file.txt -c chrY -o /pico/scratch/
usertrain/your_username/SV/2chimps_pindel
```

Parameters used to run Pindel:

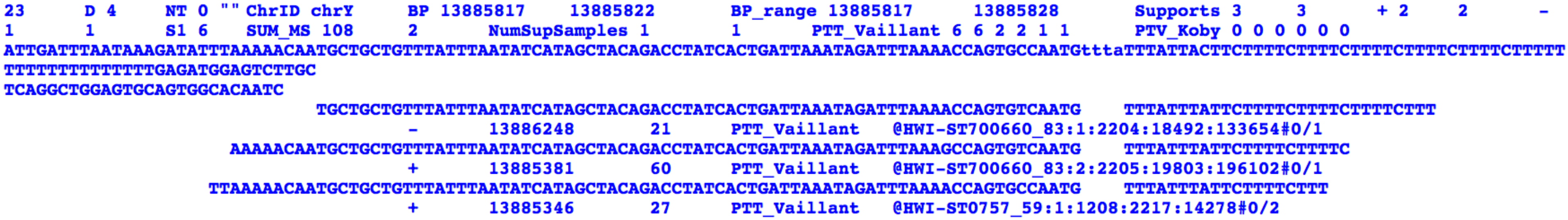
-f	The reference genome sequences in fasta format
-i	The bam config file
-c	Which chr/fragment
-o	Output prefix

Pindel will output calls for the following SVs:

- D = deletion
- SI = short insertion
- INV = inversion
- TD = tandem duplication
- LI = large insertion
- BP = unassigned breakpoints

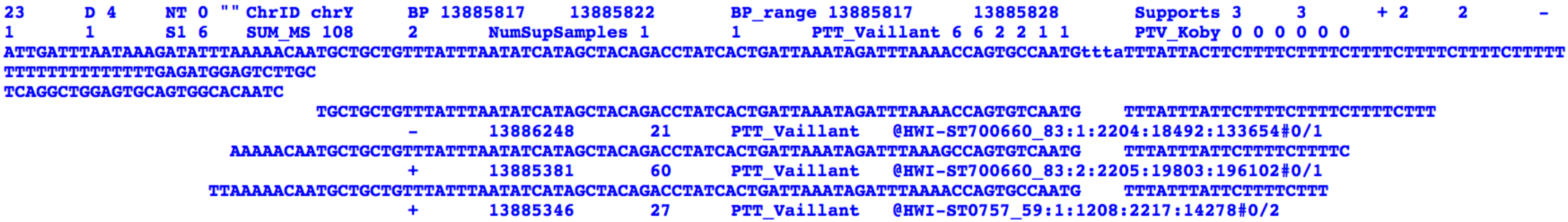
Look at the Pindel raw output:

```
cd /pico/scratch/usertrain/your_username/SV/
ls
more 2chimps_pindel_D
```



```
23      D 4     NT 0 "" ChrID chrY      BP 13885817     13885822        BP_range 13885817       13885828        Supports 3      3       + 2     2       -
1       1       S1 6    SUM_MS 108      2       NumSupSamples 1         1       PTT_Vaillant 6 6 2 2 1 1       PTV_Koby 0 0 0 0 0 0
ATTGATTTAATAAAGATATTTAAAAACAATGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAACCAGTGCCAATGtttaTTTATTACTTCTTTTCTTTTCTTTTCTTTTCTTTTT
TTTTTTTTTTTTTTGAGATGGAGTCTTGC
TCAGGCTGGAGTGCAGTGGCACAATC
                              TGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAACCAGTGTCAATG     TTTATTTATTCTTTTCTTTTCTTTTCTTT
                                              -         13886248        21      PTT_Vaillant    @HWI-ST700660_83:1:2204:18492:133654#0/1
                      AAAAACAATGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAGCCAGTGTCAATG     TTTATTTATTCTTTTCTTTTC
                                              +         13885381        60      PTT_Vaillant    @HWI-ST700660_83:2:2205:19803:196102#0/1
                    TTAAAAACAATGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAACCAGTGCCAATG     TTTATTTATTCTTTTCTTT
                                              +         13885346        27      PTT_Vaillant    @HWI-ST0757_59:1:1208:2217:14278#0/2
```

1 Index
2 The type of indel/SV
3 The length of the SV
4 "NT" indicate that the next number is the length of non-template sequences inserted
5 the length(s) of the NT fragment(s)
6 the sequence(s) of the NT fragment(s)
7-8 the identifier of the chromosome the read was found on
9,10,11 BP: the start and end positions of the SV
12,13,14 BP_range if the exact position of the SV is unclear
15 "Supports"
16 The number of reads supporting the SV
17 The number of unique reads supporting the SV (without duplicate reads)
18 +: supports from reads whose anchors are upstream of the SV
19,20 total and unique number of supporting reads whose anchors are upstream of the SV
21 -: supports from reads whose anchors are downstream of the SV
22,23 total and unique number of supporting reads whose anchors are downstream of the SV
24,25 S1: a simple score, ("# +" + 1)* ("# -" + 1)
26,27 SUM_MS: sum of mapping qualities of anchor reads
28 the number of different samples scanned
29,30,31 NumSupSamples?: the number of samples supporting the SV, as well as the number of samples having unique reads supporting the SV
32+ Per sample: total nr. of supporting reads with anchors upstream, total nr. of unique supporting reads with anchors upstream, total nr. of supporting reads with anchors downstream, total nr. of unique supporting reads with anchors downstream.



```
23      D 4     NT 0 "" ChrID chrY      BP 13885817     13885822        BP_range 13885817       13885828        Supports 3      3       + 2     2       -
1       1       S1 6    SUM_MS 108      2       NumSupSamples 1 1       PTT_Vaillant 6 6 2 2 1 1        PTV_Koby 0 0 0 0 0 0
ATTGATTTAATAAAGATATTTAAAAACAATGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAACCAGTGCCAATGtttaTTTATTACTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTT
TTTTTTTTTTTTTTGAGATGGAGTCTTGC
TCAGGCTGGAGTGCAGTGGCACAATC
                             TGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAACCAGTGTCAATG    TTTATTTATTCTTTTCTTTTCTTTTCTTT
                                      -      13886248       21      PTT_Vaillant   @HWI-ST700660_83:1:2204:18492:133654#0/1
                     AAAAACAATGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAGCCAGTGTCAATG    TTTATTTATTCTTTTCTTTTC
                                      +      13885381       60      PTT_Vaillant   @HWI-ST700660_83:2:2205:19803:196102#0/1
                   TTAAAAACAATGCTGCTGTTTATTTAATATCATAGCTACAGACCTATCACTGATTAAATAGATTTAAAACCAGTGCCAATG    TTTATTTATTCTTTTCTTT
                                      +      13885346       27      PTT_Vaillant   @HWI-ST0757_59:1:1208:2217:14278#0/2
```

Output file format:

- Reference sequence
- Sequence of the read
- Whether the anchor read is upstream(+) or downstream(-)
- The position of the mapped half of the paired-end read
- Mapping quality of the mapped read
- Sample name
- Read ID

Browsing thought the deletions output, which seems to be the most common lenght of these events? Which is the longest one you can find?

How many deletions and insertions in total have been called from these samples?

(hint: try the “tail” command)

Use pindel2vcf (conversion of Pindel output to VCF format and some filtering of SVs):

For full information about pindel2vcf options do (or check the handbook):

```
pindel2vcf
```

Then convert and filter our small insertion calls.

```
pindel2vcf -r /pico/scratch/usertrain/your_username/SV/
ucsc.hg19.fasta -R hg19 -d Feb2009 -p /pico/scratch/usertrain/
your_username/SV/2chimps_pindel_SI -v /pico/scratch/usertrain/
your_username/SV/2chimps_pindel_SI.vcf -c chrY -is 5 -b -e 5 -sr
13870437 -er 16095787
```

Parameters used to run pindel2vcf:

-r The reference genome sequences in fasta format

-R The name and version of the reference genome

-d The date of the version of the reference genome used

-p The name of the pindel output file containing the SVs

-v The name of the output vcf-file

-c The name of the chromosome

-is The minimum size of events to be reported

-b Only report events that are detected on both strands

-e The minimum number of supporting reads to report a SV

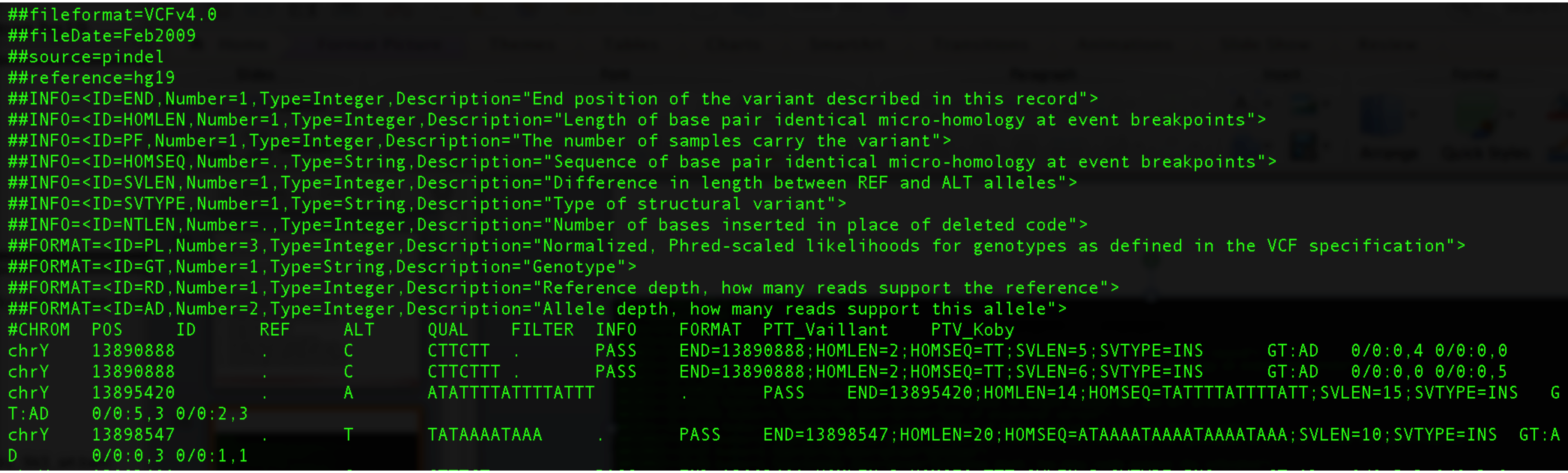
-sr The start of the region of which events are to be reported

-er The end of the region of which events are to be reported

Choice of filters depend on what you are looking for!

pindel results in vcf format

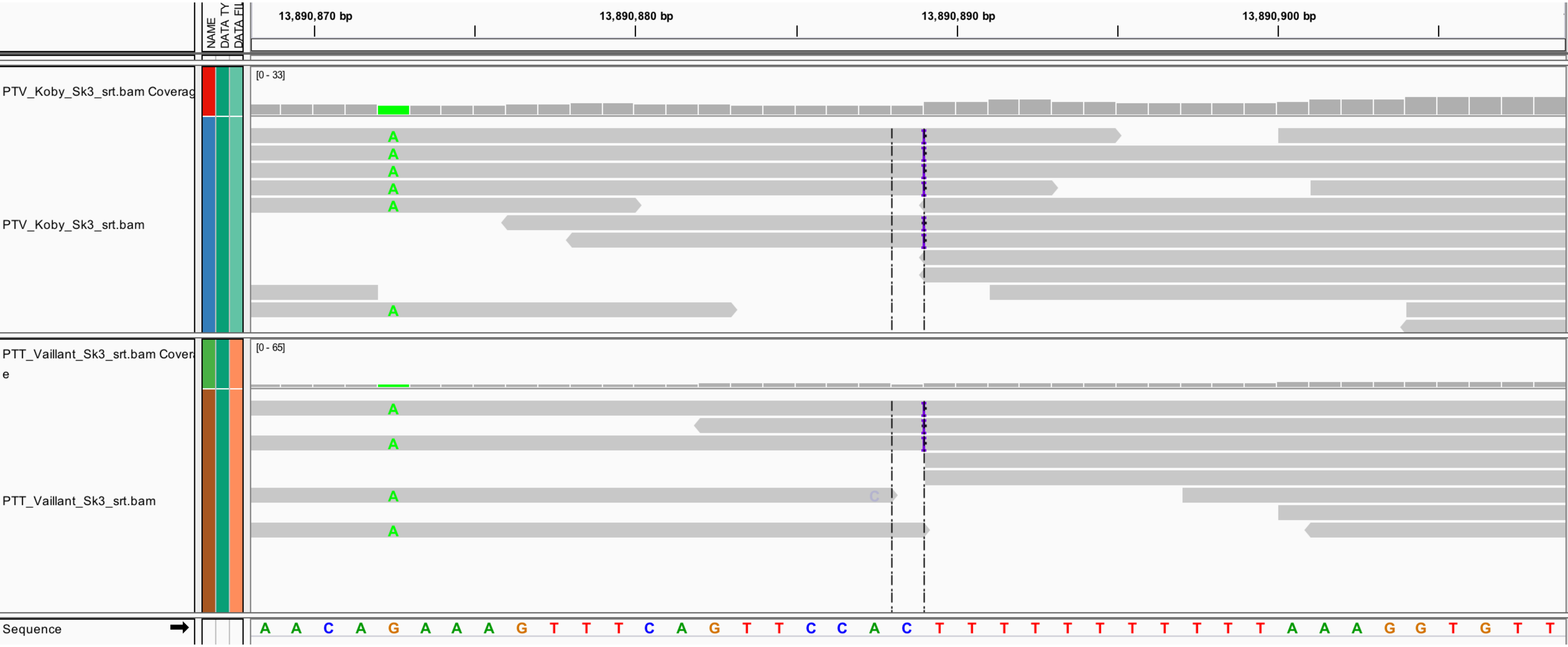
```
more 2chimps_pindel_SI.vcf
```



```
##fileformat=VCFv4.0
##fileDate=Feb2009
##source=pindel
##reference=hg19
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=1,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=PF,Number=1,Type=Integer,Description="The number of samples carry the variant">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=NTLEN,Number=.,Type=Integer,Description="Number of bases inserted in place of deleted code">
##FORMAT=<ID=PL,Number=3,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=RD,Number=1,Type=Integer,Description="Reference depth, how many reads support the reference">
##FORMAT=<ID=AD,Number=2,Type=Integer,Description="Allele depth, how many reads support this allele">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  PTT_Vaillant    PTV_Koby
chrY    13890888        .       C       CTTCTT  .       PASS    END=13890888;HOMLEN=2;HOMSEQ=TT;SVLEN=5;SVTYPE=INS      GT:AD   0/0:0,4 0/0:0,0
chrY    13890888        .       C       CTTCTTT .       PASS    END=13890888;HOMLEN=2;HOMSEQ=TT;SVLEN=6;SVTYPE=INS      GT:AD   0/0:0,0 0/0:0,5
chrY    13895420        .       A       ATATTTTATTTTATTT        .       PASS    END=13895420;HOMLEN=14;HOMSEQ=TATTTTATTTTATT;SVLEN=15;SVTYPE=INS   G
T:AD    0/0:5,3 0/0:2,3
chrY    13898547        .       T       TATAAAATAAA     .       PASS    END=13898547;HOMLEN=20;HOMSEQ=ATAAAATAAAATAAAATAAA;SVLEN=10;SVTYPE=INS  GT:A
D       0/0:0,3 0/0:1,1
```

How many small insertion calls were in the raw file and how many are in the filtered vcf?

```
                                                              PTT_Vaillant    PTV_Koby
chrY 13890888      .      C      CTTCTT    .      PASS
     END=13890888;HOMLEN=2;HOMSEQ=TT;SVLEN=5;SVTYPE=INS GT:AD      0/0:0,4
     0/0:0,0
chrY 13890888      .      C      CTTCTTT  .      PASS
     END=13890888;HOMLEN=2;HOMSEQ=TT;SVLEN=6;SVTYPE=INS GT:AD      0/0:0,0
     0/0:0,5
```



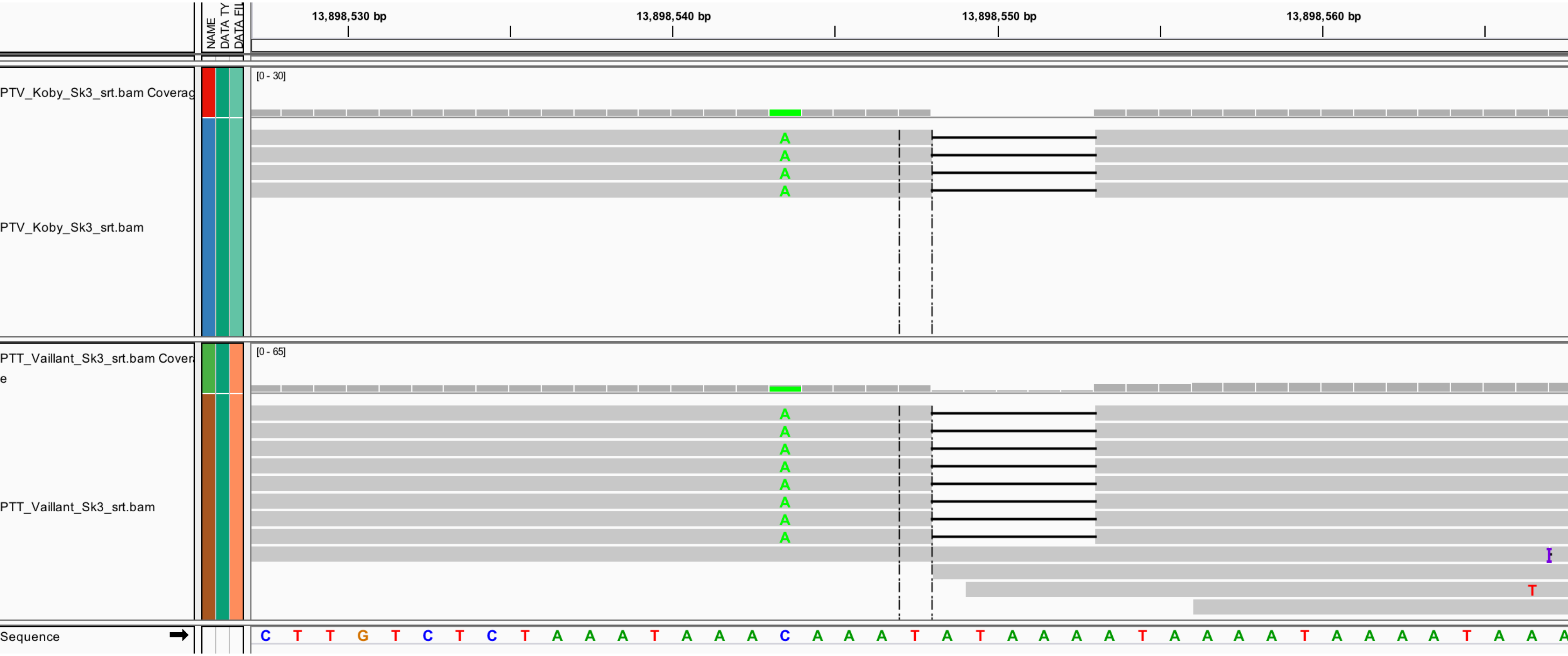
Use pindel2vcf:

Convert and filter our deletion calls:

```
pindel2vcf -r /pico/scratch/usertrain/your_username/SV/
ucsc.hg19.fasta -R hg19 -d Feb2009 -p /pico/scratch/usertrain/
your_username/SV/2chimps_pindel_D -v /pico/scratch/usertrain/
your_username/SV/2chimps_pindel_D.vcf -c chrY -is 3 -b -e 4 -sr
13870437 -er 16095787
```

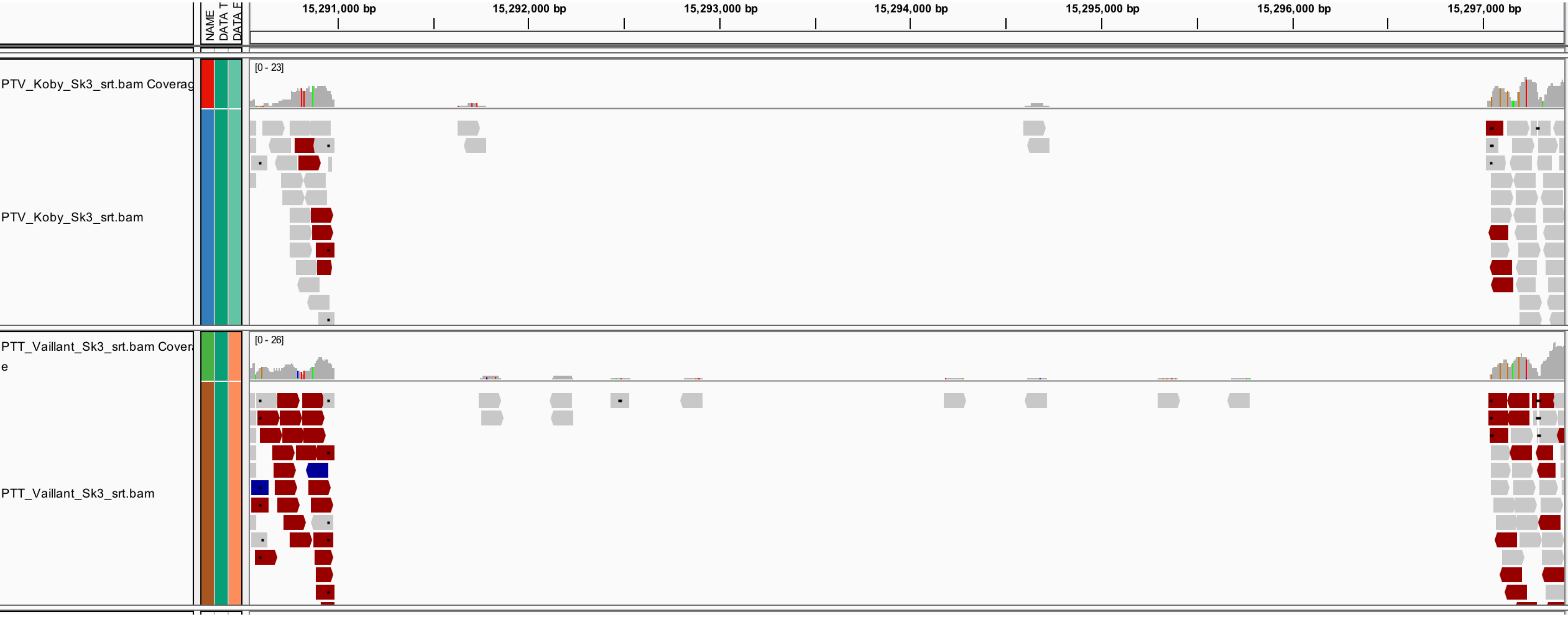
How many deletion calls were in the raw file and how many are in the filtered vcf?

```
                                                              PTT_Vaillant    PTV_Koby
chrY 13898547      .      TATAAA   T      .      PASS
      END=13898552;HOMLEN=15;HOMSEQ=ATAAAATAAAATAAA;SVLEN=-5;SVTYPE=DEL
      GT:AD      1/1:1,10   0/0:1,4
```



PTT_Vaillant PTV_Koby

chrY 15290968 . T n(6078) TGAGATGCAGTCTTGCTCTGTTGCCCAGATTGGAG .
PASS END=15297046;HOMLEN=0;SVLEN=-6078;SVTYPE=RPL;NTLEN=34
GT:AD 0/0:0,8 0/0:0,1



Choose a SV calling and filter them using different parameters.

Check the calls on IGV to get a feel of how the mapping of those positions looks.
How reliable things look to you?